Semantic-Aware Real-Time Correlation Tracking Framework for UAV Videos

Xue, Xizhe ; Li, Ying; Yin, Xiaoyue ; Shang, Changjing; Peng, Taoxin; Shen, Qiang

Published in:
IEEE Transactions on Cybernetics

DOI:
10.1109/TCYB.2020.3005453

Publication date:
2022

Citation for published version (APA):
Xue, X., Li, Y., Yin, X., Shang, C., Peng, T., & Shen, Q. (2022). Semantic-Aware Real-Time Correlation Tracking Framework for UAV Videos. *IEEE Transactions on Cybernetics*, *52*(4), 2418-2429. https://doi.org/10.1109/TCYB.2020.3005453

Semantic-Aware Real-Time Correlation Tracking Framework for UAV Videos

Xizhe Xue, Ying Li, Xiaoyue Yin, Changjing Shang, Taoxin Peng, and Qiang Shen

***Abstract*—Discriminative correlation filter (DCF) has contributed tremendously to address the problem of object tracking benefitting from its high computational efficiency. However, it has suffered from performance degradation in unmanned aerial vehicle (UAV) tracking. This article presents a novel semantic-aware real-time correlation tracking framework (SARCT) for UAV videos to enhance the performance of DCF trackers without incurring excessive computing cost. Specifically, SARCT first constructs an additional detection module to generate ROI proposals and to filter any response regarding the target irrelevant area. Then, a novel semantic segmentation module based on semantic template generation and semantic coefficient prediction is further introduced to capture semantic information, which can provide precise ROI mask, thereby effectively suppressing background interference in the ROI proposals. By sharing features and specific network layers for object detection and semantic segmentation, SARCT reduces parameter redundancy to attain sufficient speed for real-time applications. Systematic experiments are conducted on three typical aerial datasets in order to evaluate the performance of the proposed SARCT. The results demonstrate that SARCT is able to improve the accuracy of conventional DCF-based trackers significantly, outperforming state-of-the-art deep trackers.**

***Index Terms*—Detection proposals, discriminative correlation filter (DCF), semantic information, unmanned aerial vehicle (UAV) tracking.**

I. Introduction

SIGNIFICANT developments in unmanned aerial vehicles (UAVs) have been witnessed recently, delivering more diversity and flexibility for photography than common surveillance cameras with fixed camera angles, scale, and view [1]. Aerial object tracking [2] has enabled many new and important applications in computer vision [3], such as crowd monitoring, target following, and aerial navigation. Aiming to analyze the movement of a certain target, visual tracking algorithms [4]–[6] locate their bounding boxes on the video stream according to the given initial state in the first frame. Although methods based on the correlation filter [4] and Siamese network [5], [6] have achieved decent performance in general scenes, robust and accurate aerial tracking [2], [7] remains active research because of the particular challenging factors, such as unpredictable weather conditions, changing flying altitude, and shaking camera views. To be more specific, aerial objects are usually tiny and move very fast or rotate drastically, resulting in difficulties in tracking them. In addition, apart from the interference of shadows and background introduced by high incline shots, aerial videos captured at poor light conditions are likely to lose the otherwise abundant texture information and sharp details [1].

Real-time processing on aerial videos is a prerequisite for practical applications. From this viewpoint, discriminative correlation filter (DCF)-based trackers [4], [8] are focused on for their strengths on both accuracy and speed. These kinds of methods always employs the fast Fourier transform (FFT) to convert the models from the time domain to the frequency domain, in which the convolution operation is transformed into multiplication, greatly reducing the computational cost during the process of locating a target. However, targets captured in UVA videos are often obscure in shadows and fully or partly occluded by other objects, for example, trees, roofs, and signs [9]. Under such circumstances, traditional DCF algorithms still assume that the point with the highest value in the response map represents the target location. If a template is contaminated with shielding, the tracker may be misled and cannot relocate the target when it is lost for a period and appears again. Therefore, it is necessary to introduce a detection module [10], [11] to help the tracker recover from common challenges in aerial videos, such as temporary or persistent occlusions.

While tracking-by-detection methods coarsely improve the confidence of detection proposals, how to repress no-target object regions in a bounding box and finely locate the target

Manuscript received December 18, 2019; revised March 29, 2020; accepted June 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61871460; in part by the Shaanxi Provincial Key Research and Development Program under Grant 2020KW-003; in part by the Fundamental Research Funds for the Central Universities under Grant 3102019ghxm016; in part by the Innovation Foundation for Graduate Students School-Enterprise Cooperation of Northwestern Polytechnical University under Grant XQ201905; and in part by the Sêr Cymru II Strategic Partner Acceleration Award Programme, U.K., under Grant 80761-AU201. This article was recommended by Associate Editor Q. Meng. *(Corresponding author: Ying Li.)*

Xizhe Xue, Ying Li, and Xiaoyue Yin are with the School of Computer Science and Engineering, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xuexizhe@mail.nwpu.edu.cn; lybyp@nwpu.edu.cn; 2015302412@mail.nwpu.edu.cn).

Changjing Shang and Qiang Shen are with the Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3DB, U.K. (e-mail: cns@aber.ac.uk; qqs@aber.ac.uk).

Taoxin Peng is with the School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, U.K. (e-mail: t.peng@napier.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.3005453

remains a challenging question in aerial tracking. Because DCF trackers locate a target by correlation operations, they are easily disturbed when the appearance of the target changes dramatically within the same background. To distinguish pixels belonging to the target from the background, semantic segmentation methods that can classify images at pixel level have been taken into consideration. Semantics has been commonly used as *a priori* information for tracking methods that deal with a specific type of target, such as human tracking [12] and vehicle tracking [13]. However, the categories of targets vary and are not provided in the generic visual tracking as well as aerial tracking. Under this circumstance, semantic information [14] is difficult to apply directly.

To handle the aforementioned issues, we propose a semantic-aware real-time aerial object tracking framework in this article. Specifically, two additional enhancements: one corresponding to the detection module and another to the semantic segmentation module, are introduced to modify the traditional DCF-based trackers. The resulting tracking framework is able to cope with serious occlusion and deformation in aerial videos while retaining their characteristic speed and real-time capability. Furthermore, the proposed approach helps suppress noise from the background, thereby locating the target more precisely supported by the rich semantic information obtained from the ingenious semantic segmentation module.

The key contributions of this work are summarized as follows.

1) To improve the tracking accuracy on videos captured by small UAVs, a novel aerial tracking framework capable of effectively locating targets and suppressing interference from the background is proposed and tested on local servers. Based on the conventional DCF tracker, additional detection and semantic segmentation modules using deep networks are introduced into the proposed framework to enhance the aerial tracking performance. Through sharing deep features and specific network layers, the number of model parameters is reduced greatly and real-time performance for target tracking on UVA videos is ensured.
2) The background suppression problem in aerial tracking is analyzed and an efficient semantic segmentation module is designed to resolve this problem. The semantic templates are generated through fully convolutional networks (FCNs), and the semantic coefficients are computed from the prediction head. Weighing the semantic templates with their corresponding coefficients, we can obtain the ROI masks to help locate targets on UAV videos precisely. Different from other semantic-aware tracking methods, which are merely trained offline, the proposed approach is committed to initially train the module offline and update it online.
3) Extensive experimental results with four representative DCF baselines on three typical aerial tracking datasets are reported to present the advantages of the proposed aerial tracking framework. Both handcrafted feature-based trackers and state-of-the-art deep leaning tracking methods have been experimentally compared with the proposed framework, which performs favorably against the other trackers according to systematic evaluations.

The remainder of this article is organized as follows. Some related works are reviewed in Section II. We introduce the proposed aerial tracking framework in Section III while experimental results as well as the analyses are offered in Section IV. Besides, some conclusions are summarized and future research proposals are suggested in Section V.

II. Related Work

A. DCF-Based Tracker

Since Bolme *et al.* [15] first explored a minimum output sum of squared error (MOSSE) filter to visual tracking field, DCF tracking methods have been widely researched and welcomed for their high computational efficiency. By learning a correlation filter, a typical DCF tracker computes the circular convolution response and generates a spatial confidence map (also called a response map), where the position with the maximum response indicates the location of the target.

Following this general approach, various DCF algorithms have been developed rapidly over the last few years. By introducing the cycle shift and kernel trick, Henriques *et al.* [16] designed a kernelized correlation filter (KCF) with multichannel HOG features to achieve a high-speed tracking of 172 frames per second (FPS). Except for the traditional translation filter, the DSST tracker [17] employs an additional scale filter in order to detect the scale changes of targets. Fusing the HOG and color name features, Bertinetto *et al.* [18] proposed a Staple algorithm with real-time running speed so that deformations and color changes of targets can be handled well. While a number of such methods exist, the boundary effects caused by the cycle shift still trouble researchers. With the help of a spatial regularization term, the SRDCF tracker [19] alleviates this problem by punishing the filter coefficients in the boundary region. Although the handcrafted features [15]–[19] are widely utilized, the application of deep features [20], [21] is also investigated for more accurate and comprehensive appearance representation. Combining the handcrafted features and deep features, the resultant algorithms [22] are able to describe the object at different levels of abstraction, therefore, contributing to obtaining better tracking performance. Indeed, even DCF trackers have achieved the state-of-the-art performance in multiple aerial tracking datasets [1], [9], [23], potential promotion space still exists.

B. Tracking by Detection

Aerial tracking is up against some similar challenges to long-term tracking, where tracking by detection has been widely studied extensively. Tracking-by-detection methods always learn initial discriminative models (e.g., a support vector machine, SVM) to detect the targets. One of the representative algorithms among these is TLD [24], which decomposes the entire visual tracking task into three subtasks, namely, tracking, learning, and detection. Each subprocess reinforces one another. In particular, TLD utilizes the current optical-flow information to predict the position of the target

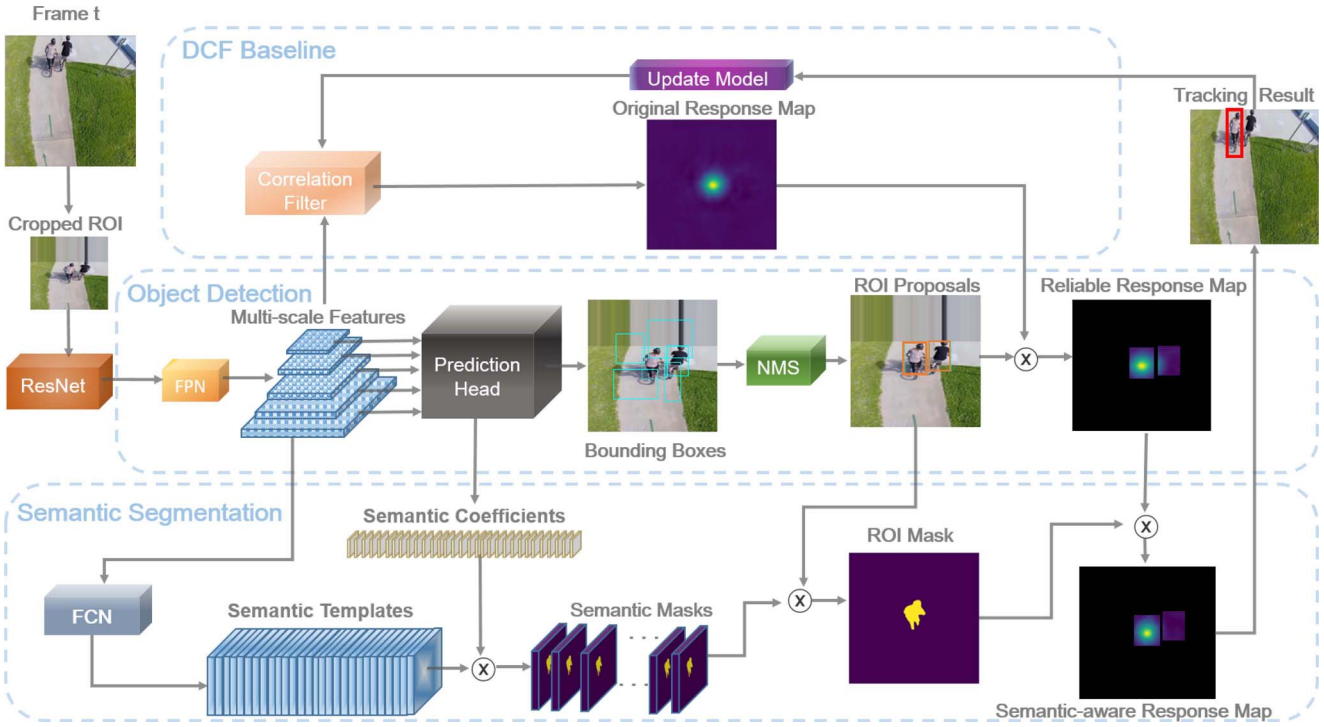


Fig. 1. Proposed aerial tracking framework.

in the next frame. Meanwhile, the detection module generates a great number of proposals, in which the one accepted by most filters is taken as the final detection result. If the tracking module fails, the detection procedure will reinitialize the tracker, helping it recomplete the task. This kind of discrimination is usually based on the assumption that the target's motion and shape change smoothly, but this assumption is often too idealized to happen in the real world. Inspired by this observation, a multientropy minimization (MEEM) tracker [25] has been presented to tackle the underlying problem by maintaining a collection of snapshots and choosing the best prediction from them. Also, Ma *et al.* [26] have proposed a DCF-based method, where a k-nearest neighbor classifier is utilized to collect training samples and an online random ferns classifier is employed to redetect the target.

C. Semantic Segmentation

Semantic segmentation is a basic task in computer vision, by which the specific regions of an image are labeled according to what is being shown. Recently, the most attractive deep architecture for semantic segmentation is FCN [27]–[31]. Combining detailed information from a shallow layer with coarse information from the deep one, Long *et al.* [27] first realized accurate segmentation by a convolutional network. However, earlier FCN-based methods suffered from low-resolution prediction generally. Aiming to overcome this limitation, U-Net [28] is proposed, in which a contracting and symmetric expanding path is introduced to capture the context, achieving precise localization before learning a successive convolution layer. Also, in SegNet [29], certain pooling indices are saved in the max-pooling step of the novel encoder. Then, these indices are taken to help nonlinear upsampling, which forms the sparse upsampled maps. Relying on them and the trainable convolutional filters, dense and precise feature maps can be produced. Furthermore, to enable efficient high-resolution prediction, Lin *et al.* [30] proposed RefineNet, where features of different scales are fused during the downsampling process and the rich background contexts are retained by chained residual pooling. To handle the problem of segmenting objects at multiple scales, Chen *et al.* designed Deeplabv3+ [31] by employing atrous convolution layers with multiple atrous rates in cascade or in parallel.

III. Proposed Method

The flowchart of the proposed aerial tracking framework is depicted in Fig. 1, which consists of three modules: 1) DCF tracking baseline; 2) object detection; and 3) semantic segmentation. As is shown, the ROI region is first cropped around the center of the target in the previous frame. Then, this image block is fed into a backbone network (such as Resnet [32]) with a feature pyramid network (FPN) [33] to extract multiscale deep features, forming the inputs to the three modules. Specifically, the resulting multiscale deep features are first utilized (with any typical DCF-based tracker, for example, KCF [16], DSST [17], Staple [18], and ECO [22]) to generate an original response map. Next, the detection module exploits the deep features to produce ROI proposals. The original response map is fused with the bounding boxes of the ROI proposals to form a reliable response map with less background interference. Meanwhile, the semantic segmentation network processes the deep features to acquire semantic templates and coefficients, which jointly make up the required

TABLE I
STRUCTURE OF THE SEMANTIC TEMPLATES GENERATION BRANCH

Layer	1	2	3	4	5	6
Filter	$256 \times conv3$	$256 \times conv3$	$256 \times conv3$	$upsample \times 2$	$256 \times conv3$	$32 \times conv1$

semantic masks. After choosing the ROI mask from tons of the computed semantic masks and combining the reliable response map with it, the sematic-aware response map can be calculated, in which the position with the maximum response value offers a hint of the target's location.

Note that both the detection and semantic segmentation modules in the proposed tracking framework are class agnostic, that is, they are not reliant on any prior knowledge of the target. These two modules aim to find the most similar one with what has been given in the first frame rather than assume a class-specific target. Even when similar and reliable semantics have not been mined, the proposed approach can still perform robust aerial tracking, solely relying on the correlation filter and detection module.

A. DCF Baseline

Considered as a classifier, DCF-based trackers are always trained by minimizing the least-square errors between the training samples x_i and the templates y_i through online training

$$\min_{w} \sum_{i} L(f(w, x_i), y_i) + \lambda \|w\|^2 \tag{1}$$

where $f(\cdot)$ and w denote the relation sought and model parameters, respectively. $L(\cdot)$ is the l_2 normal loss function. Notably, to prevent overfitting, a regularization parameter λ is introduced.

After being converted into the frequency domain, the trained correlation filter on the h-th($h \in \{1, \ldots, H\}$) dimension is, thus, as follows:

$$W^h = \frac{Y \odot \overline{X^h}}{\sum_{i=1}^{H} X^i \odot \overline{X^i} + \lambda}. \tag{2}$$

The bar represents complex conjugation and the operator $\odot$ denotes the elementwise product. When the feature vector z of an image patch is obtained, the response map in the next frame can be generalized as

$$r = F^{-1}\left(\sum_{h=1}^{H} W^h \odot \overline{Z^h}\right) \tag{3}$$

where the operator F^{-1} denotes the inverse FFT. The target state is then estimated by finding the peak of the response map r.

B. Object Detection

First, the target is coarsely located with high efficiency based on a DCF-based tracker, which maintains the temporal consistency of neighboring frames. Then, a detection module is drawn into the proposed framework to refine the initial position computed from the original response map. In specific, our detection approach is a novel CNN, which generates a fixed-size collection of bounding boxes. By applying a nonmaximum suppression (NMS) operation to them, we select ROI proposals according to the scores of class instances in those boxes.

In the proposed semantic-aware real-time correlation tracking (SARCT) framework, ResNet-50 with FPN is considered as the default backbone to obtain different layer features. Specifically, the output of each stage's last residual block in ResNet is used, and the resulting features $\{\text{conv3}, \text{conv4}, \text{conv5}\}$ are denoted as $\{C3, C4, C5\}$, which have strides of $\{8, 16, 32\}$ pixels, respectively. Similar to RetinaNet [34], feature pyramid levels $P3$–$P7$ have been applied in this framework, in which $P3$–$P5$ are obtained from the output of the corresponding ResNet residual stage using top-down and lateral connections. The $P6$ and $P7$ levels are created by simply applying stride 2 and stride 4 with max-pooling to $C5$. The feature dimensionality is fixed as 256 in this work to ensure that all levels of the pyramid are able to share the same classifier.

In the implementation, the prediction head is constructed with a 3×3 convolutional layer followed by two siblings 3×3 convolutions to carry out the classification and bounding box regression. A head of the identical structure (3×3 conv and two siblings 3×3 convs) is attached to each level on the generated feature pyramid. Anchors representing a group of reference boxes are defined of various scales and aspect ratios to deal with the targets with different shapes. Depending on them, it is convenient to perform the object/nonobject criterion and target box regression. It is worth mentioning that the anchors are specified to have areas of $\{24, 48, 96, 192, 384\}$ pixels on $\{P3, P4, P5, P6, P7\}$ in the proposed framework, respectively. The anchors' multiple ratios are set to $\{1:2, 1:1, 2:1\}$ at each level.

Having obtained the ROI proposals from the detector, the original response map is refined to becoming a reliable response map R_r. By retaining the response in the limited ROI proposal areas, the reliable response map suppresses the background disturbance and improves the performance of the resulting aerial tracking algorithm.

C. Semantic Segmentation

To achieve a more precise location on aerial videos, an adaptive semantic segmentation module is proposed and drawn into our framework. It digs out the boundaries' information of object proposals flexibly through FCN-based semantic templates generation and semantic coefficient prediction. Compared with traditional semantic segmentation models that are completely dependent on the offline training, our method can learn abundant feature representation in offline training and then adapt to appearance changes of the target on aerial videos through online updating.

TABLE II
STRUCTURE OF THE INTEGRATED PREDICTION HEAD (N_a REPRESENTS THE NUMBER OF ANCHORS)

Layer1		$256 \times conv3$	
	Box Regression branch	Classification branch	Semantic coefficient Prediction branch
Layer2	$4 \times N_a \times conv3$	$c \times N_a \times conv3$	$k \times N_a \times conv3$

Semantic Template Generation: The semantic template generation branch predicts a set of k ($k = 32$) pixel-level semantic templates for an input patch. Table I shows the structure of this branch, in which conv3 and conv1 denote the convolution kernel with size 3×3 and 1×1, respectively. It is implemented by constructing a subnet similar to FCN [27] whose last layer has k channels and attaching it to a backbone feature layer. The deepest features in FPN are used and upsampled to one fourth the dimensionality of the input image to produce more robust masks and hence, to enhance the performance on small objects.

For the cropped ROI, its semantic templates D are extracted by this branch and vectorized as described above. Formally, this can be denoted by

$$D = \left[d_1, d_2, \ldots, d_k\right] \in R^{k \times (m \times n)} \tag{4}$$

where k is the number of templates corresponding to each channel of the branch output, and m and n represent the height and width of the semantic templates, respectively.

Semantic Coefficient Prediction: In the proposed framework, an offline-training semantic coefficient prediction branch is added to the prediction head to predict k semantic coefficients, which shares the first convolutional layer with the detection module. The k mask coefficients are calculated regarding each template from FCN. Thus, as demonstrated in Table II, instead of producing $4 + c$ coefficients per anchor in the prediction head, $4 + c + k$ are produced.

For the output semantic coefficients C, *tanh* is applied to produce more stable outputs while considering no nonlinearity

$$C = \tan\left(\left[c_1, c_2, \ldots, c_{N_a}\right]\right) \in R^{N_a \times k}. \tag{5}$$

Semantic masks M can, therefore, be obtained by

$$M = \begin{pmatrix} \max(C \times D, 1), & \text{if } p_{i,x,y} \geq 0.5 \\ \min(C \times D, 0), & \text{if } p_{i,x,y} < 0.5 \end{pmatrix} \in R^{N_a \times m \times n} \tag{6}$$

where $p_{i,x,y}$ is the value in matrix $C \times D$, which is corresponding to the semantic information in anchor i at location (x, y).

Among all N_a semantic masks, the one corresponding to the anchor with the highest score from the classification branch is considered as the ROI mask R_{ROI}. Fused R_{ROI} with the reliable response map R_r, the final semantic-aware response map R is constructed

$$R = (1 - p)R_r + pR_{\text{ROI}} \tag{7}$$

where p denotes the semantic weight.

From the above, the new target state can be inferred by finding out the peak of the sematic-aware response map R. By introducing pixel-level semantic information adaptively and inhibiting the background noise, the result from the detection step can be further refined. A target template with a precise mask rather than a simple bounding box contributes to achieving a better tracking performance.

D. Model Offline Training and Templates Online Updating

Model Offline Training: In the present work, the smooth-$L1$ loss L_{loc} and softmax cross-entropy loss L_{cls} are utilized to train box regression and class prediction branch, respectively.

Specifically, for object class u

$$L_{\text{loc}}\left(P^u, V\right) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}\left(P_i^u - V_i\right) \tag{8}$$

where P is the bounding box regression result and V is the ground truth, and

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$
$$L_{\text{cls}}(\text{Score}, u) = -\log \text{Score}_u \tag{9}$$

in which Score_u denotes a discrete probability distribution from network output.

For the semantic segmentation module, given the labeled ground truth G, the proposed semantic templates D and coefficients C can be optimized by minimizing the cross-entropy loss L_{seg} between S labels and N_a candidates from prediction head

$$L_{\text{seg}}(C, D, G) = -\sum_{j=1}^{S} \sum_{i=1}^{N_a} G_j \log\left(C_{ij} \times D_{ij}\right). \tag{10}$$

Notably, the DCF module of the proposed SARCT does not participate in any training process, while the object detection and semantic segmentation modules are simultaneously trained on the instance segmentation part of the COCO dataset [35]. The entire offline optimization process aims to minimize the sum of the three branches' loss functions.

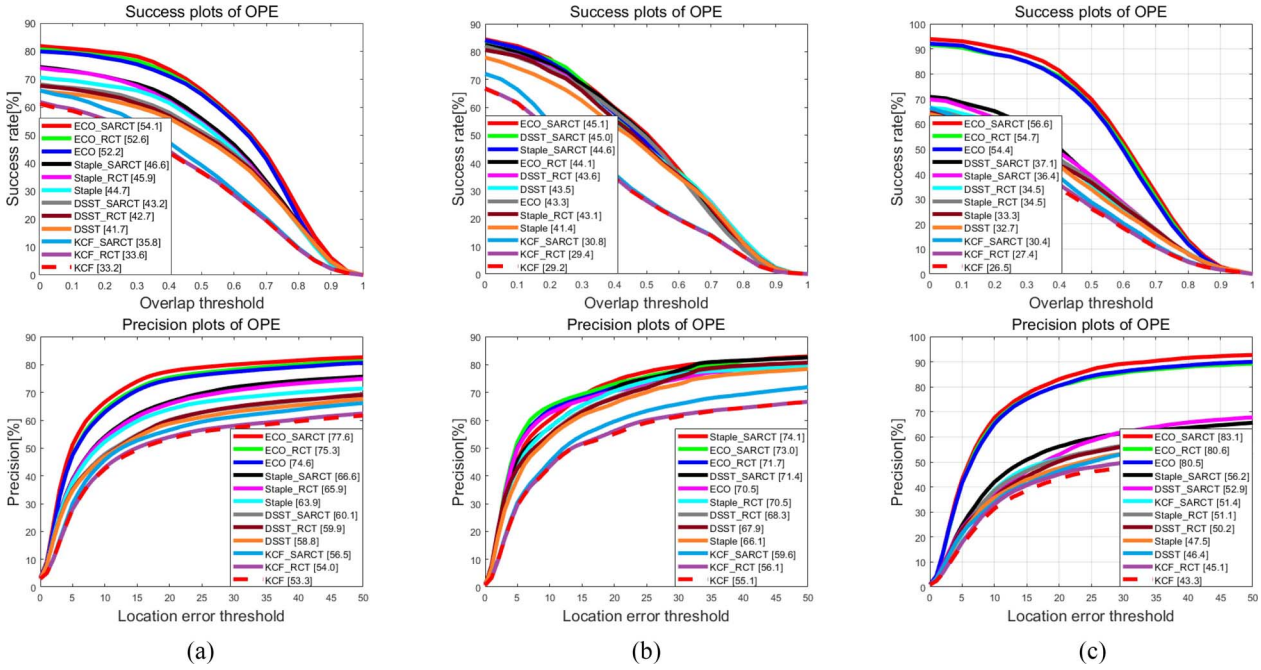
Templates Online Updating: The semantic templates are updated online with learning rate η

$$D_t = (1 - \eta)D_{t-1} + \eta D_t \tag{11}$$

where t and $t - 1$ denote the t-th and $(t - 1)$th frame, respectively.

IV. EXPERIMENTAL EVALUATION

In this section, the proposed SARCT framework is systematically evaluated on three UAV tracking benchmarks: 1) UAVDT [1]; 2) UAV123 [9]; and 3) DTB [23], which totally include 243 challenging image sequences altogether involving over 90 000 frames. Following the protocol used in recently published methods [2], [7], [22], the results in one-pass evaluation (OPE) [36] are reported. The evaluation is based on two performance metrics: 1) success plot and 2) precision plot. The success plot illustrates the ratios of successful frames over the range of thresholds [0, 1], where the area under the curve (AUC) is included. The precision plot shows the average distance precision (DP) along with a range of thresholds, and the score of average DP at 20 pixels per tracker is given.



(a) (b) (c)

Fig. 2. Success and precision plots of baselines, their SARCT and RCT counterparts on (a) UAV123, (b) UAVDT, and (c) DTB datasets, with precision and AUC explicitly marked in plots.

A. Implementation Details

The SARCT framework is implemented in Python 3.6 under the Pytorch1.0 platform. The value of p in (7) is set to 0.05 and the learning rate η in (11) to 0.90. In the training process, the image is resized to 550×550 and the network is trained on the COCO dataset for 8×10^5 epochs with SGD and a batch size of 8. The learning rate of training is initialized to 1×10^{-3} and divided by 10 every 2×10^5 epochs, using a weight decay of 5×10^{-4} and a momentum of 0.90. All experiments of each tracker are performed on a workstation with an Intel Xeon E5-2699 processor (2.30 GHz) and an NVIDIA 1080ti GPU.

B. Comparison With DCF Baselines

The present work is concerned with an aerial tracking framework that can adapt to almost all correlation-based tracking algorithms. To reflect this generality, four representative DCF trackers are chosen to conduct the experimental investigation: 1) KCF [16]; 2) DSST [17]; 3) Staple [18]; and 4) ECO [22], serving as baselines of the framework to show its efficiency and stability. Among them, KCF is the most classic kernelized correlation tracker with HOG features. DSST learns a separate scale correlation filter for precise object-scale prediction in tracking. Staple combines hog and color features in the step of feature extraction to achieve better performance. Utilizing deep features extracted by CNN, ECO possesses excellent tracking performance on a number of popular benchmarks with the assistance of factorized convolution operators. Furthermore, we also compare the proposed aerial tracking framework without a semantic segmentation module (abbreviated as RCT) to demonstrate the effect of semantic information.

Fig. 2 shows the results of all baseline trackers and those for their SARCT and RCT counterparts on three datasets. All SARCT trackers improve their respective baselines to certain extents. The gains over the four DCF baselines regarding the success and precision rates range from {1.5%, 1.3%} to {4.4%, 9.8%} on the three datasets. In particular, SARCT trackers significantly outperform their corresponding DCF baselines with traditional features (e.g., DSST, Staple, and KCF) in terms of both AUC and precision. Furthermore, the proposed framework also attains better performance in comparison with the modern DCF tracking algorithm using deep features (e.g., ECO). What cannot be ignored is that these improvements are achieved at a much lower computational cost, thereby potentially facilitating real-time applications for aerial tracking.

An ablative experiment conducted demonstrates that both detection and segmentation help improve the performance of each algorithm (though to a different extent) with semantic information contributing more to algorithm performance in most cases. It indicates the precise target mask information provided by the semantic segmentation module is more robust to background clutter (BC).

C. Comparison With State-of-the-Art Trackers

1) Experimental Results on the DTB Dataset: As a distinct approach, the implementation ECO_SARCT of the SARCT framework is taken for this comparative study, against state-of-the-art algorithms on the DTB dataset. The compared tracking methods are ARCF [2], TADT [37], LDES [38], MCCT [39], STRCF [40], SiamFC-tri [41], CSRDCF [42], and BACF [43]. Among them, BACF, ARCF, CSRDCF, LDES, MCCT, and STRCF are DCF trackers, while SiamFC-tri and TADT are based on the end-to-end CNN.

Overall Comparison: Fig. 3 shows the experimental results of ECO_SARCT and the aforementioned state-of-the-art methods. The proposed tracker, employing an efficient detection

TABLE III
AVERAGE NUMBER OF FPS OF SARCT METHODS AND DCF BASELINES ON THE DTB DATASET

Method	ECO	ECO_SARCT	Staple	Staple_SARCT	DSST	DSST_SARCT	KCF	KCF_SARCT
Speed(FPS)	28.3	24.8	46.4	27.6	30.6	26.5	59.5	29.2
GPU	Yes	Yes	No	Yes	No	Yes	No	Yes

TABLE IV
AVERAGE NUMBER OF FPS OF STATE-OF-THE-ART METHODS ON THE DTB DATASET

Method	ARCF	LEDS	MCCT	STRCF	SiamFC-tir	TADT	BACF	CSRDCF
Speed(FPS)	8.5	7.2	1.1	20.3	67.1	44.6	10.8	6.4
GPU	No	No	Yes	No	Yes	Yes	No	No

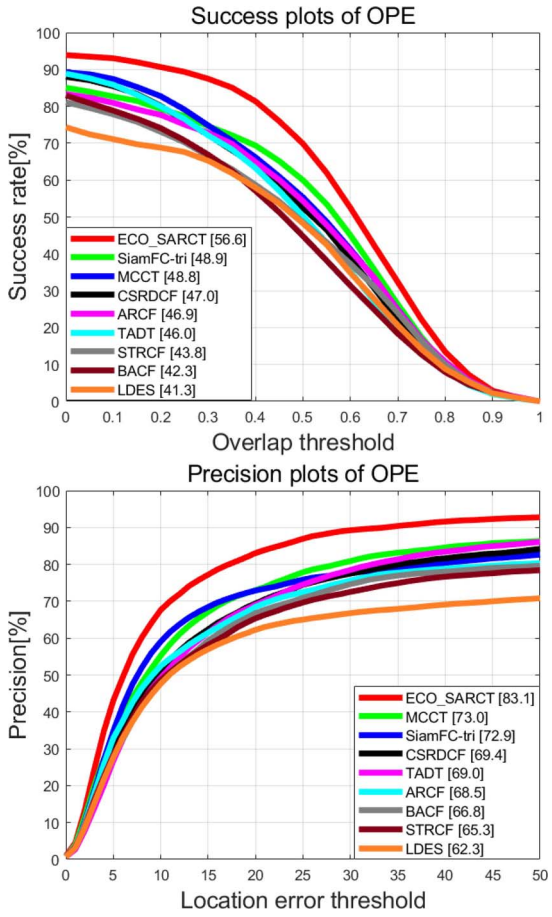


Fig. 3. Success and precision plots of ECO_SARCT and other state-of-the-art methods on the DTB dataset.

and semantic segmentation strategy for target estimation, achieves the best results on AUC and precision with scores of 56.6% and 83.1%, respectively. The proposed method significantly outperforms the second-best method SiamFC-tri or MCCT with multiple deep features by approximately 8% in AUC and 10% in precision on the DTB dataset. Compared to the aerial tracking method ARCF, ECO_SARCT still gains 9.7% and 14.6% on AUC and precision, respectively.

Attribute-Based Comparison: In this experimentation, attribute-based analyses of the SARCT framework on the DTB dataset are performed. Apart from the aspect ratio change (ARC) and BC, these aerial sequences also suffer from further difficulties, such as deformation (DEF), fast camera motion (FCM), in-plane rotation (IPR), motion blur (MB), out-of-plane rotation (OPR), occlusion (OCC), out of view (OV), similar object around (SOA), and scale variation (SV). Thus, the experiments almost cover all typical challenges involved in real-world aerial tracking problems. The success and precision plots on 11 representative attributes are demonstrated in Figs. 4 and 5. ECO_SARCT has performed favorably against other state-of-the-art trackers in all attributes defined, which fully demonstrates the effectiveness of the SARCT framework.

Speed Analysis: Tables III and IV illustrate the running speeds of SARCT methods, DCF baselines, and other state-of-the-art algorithms. Obviously, the proposed SARCT framework (which incorporates the merits of the DCF-based tracker and efficient semantic information) operates at about 25–30 FPS on a single GPU, meeting the real-time requirement.

2) Experimental Results on UAV123 and UAVDT Datasets: For verifying the robustness of the SARCT framework on aerial tracking task more comprehensively, we have performed another group of tests on UAV123 and UAVDT datasets. The comparing algorithms are the same as what has been experimented on the DTB dataset. As shown in Fig. 6, the proposed ECO_SARCT tracker has outperformed all the compared trackers based on DCF or the end-to-end CNNs on the UAV123 dataset. More specifically, ECO_SARCT (54.1%) has an advantage of 2.6% over the second-best tracker TADT (51.5%) in AUC, as well as an advantage of 4.4% and 4.9% over the second (MCCT, 73.2%) and third-best tracker (TADT, 72.7%), respectively, in terms of precision. On the UAVDT dataset, the proposed SARCT tracker (74.1%) also achieves the best performance on precision, followed by ARCF (74.0%), SiamFC-tri (73.9%), and LDES (73.9%). In addition, SiamFC-tri (47.8%) is closely followed by ECO_SARCT (45.1%) and Staple_SARCT (44.6%), in terms of AUC.

D. Qualitative Evaluations

To qualitatively compare the performance of the proposed method, Fig. 7 shows the tracking results of two baselines (ECO and Staple), their SARCT counterparts, and DCF-based aerial tracker ARCF on different aerial sequences from the typical aerial tracking datasets. Several screenshots undergoing various challenges are shown in Fig. 7, from top-down are from sequences BMX4, S0103, and bird1, demonstrating the robustness of our algorithm in complex scenarios.

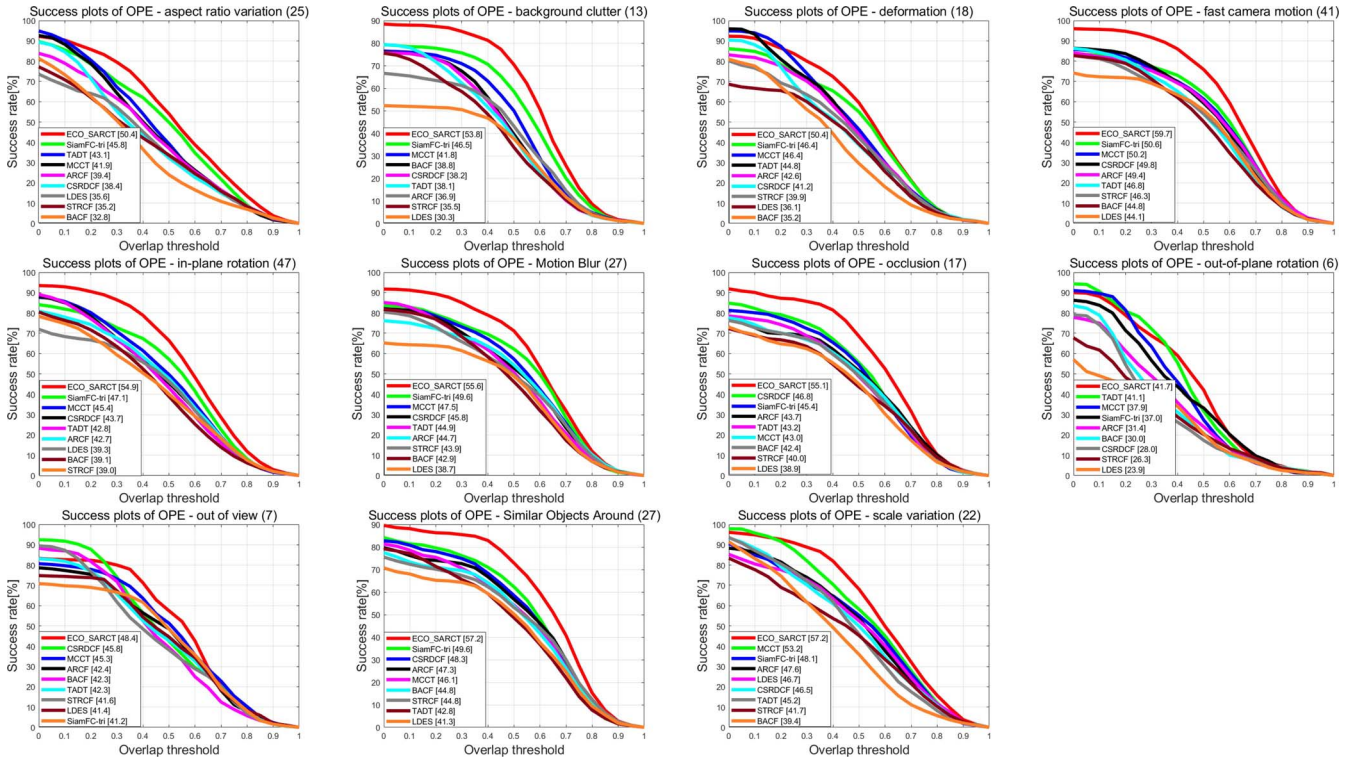


Fig. 4. Attribute-based evaluation. Success plots on 11 attributes-based comparison between ECO_SARCT and state-of-the-art trackers on the DTB dataset.

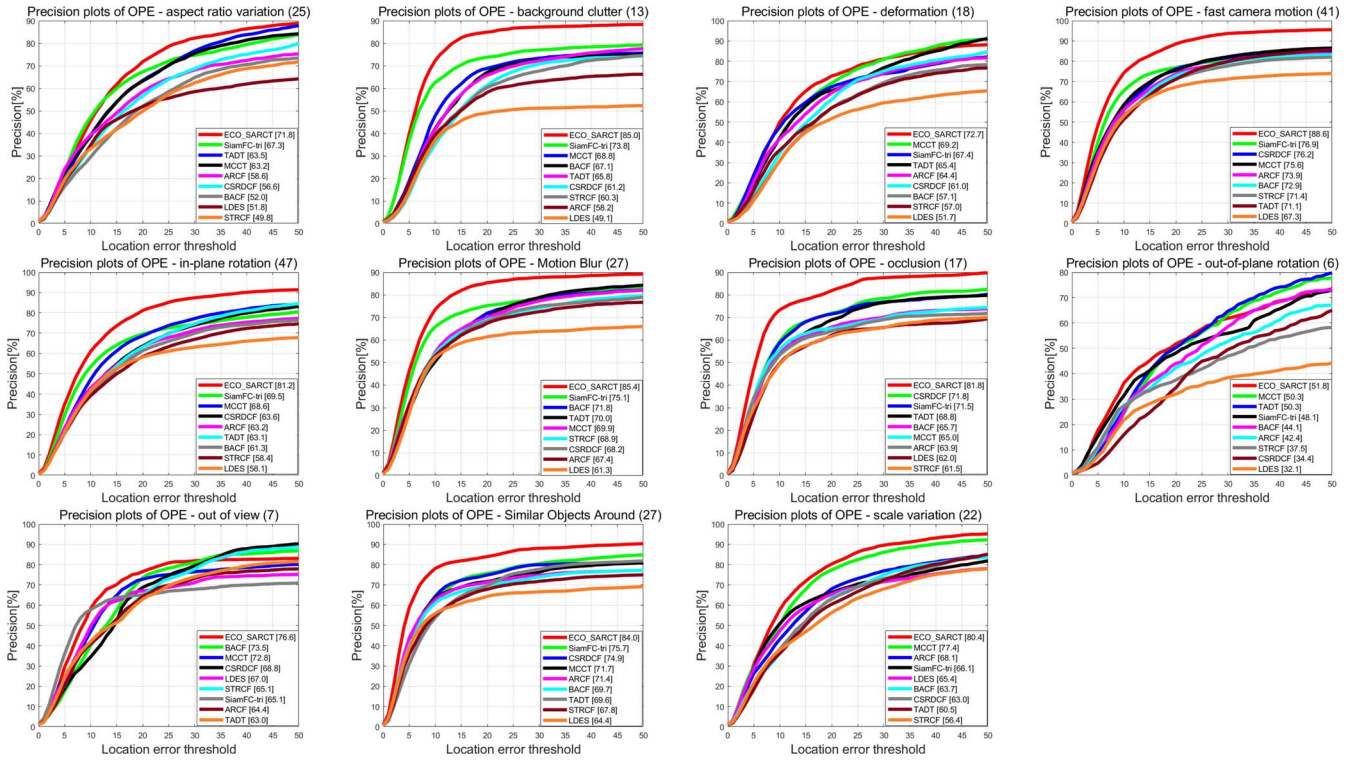


Fig. 5. Attribute-based evaluation. Precision plots on 11 attributes-based comparison between ECO_SARCT and state-of-the-art trackers on the DTB dataset.

Different extents of SVs exist on these sequences. The proposed ECO_SARCT predicts both the scale and position of the target accurately, even for the target that experiences significant appearance and scale changes on sequences BMX4 and bird1. Fig. 7(a) illustrates that benefiting from the detection and semantic segmentation modules, the proposed Staple_SARCT keeps tracking the biker for a long time while Staple lost target very early. According to these results on challenging sequence bird1, almost all other algorithms fail to locate the target because of its deformation, fast motion,

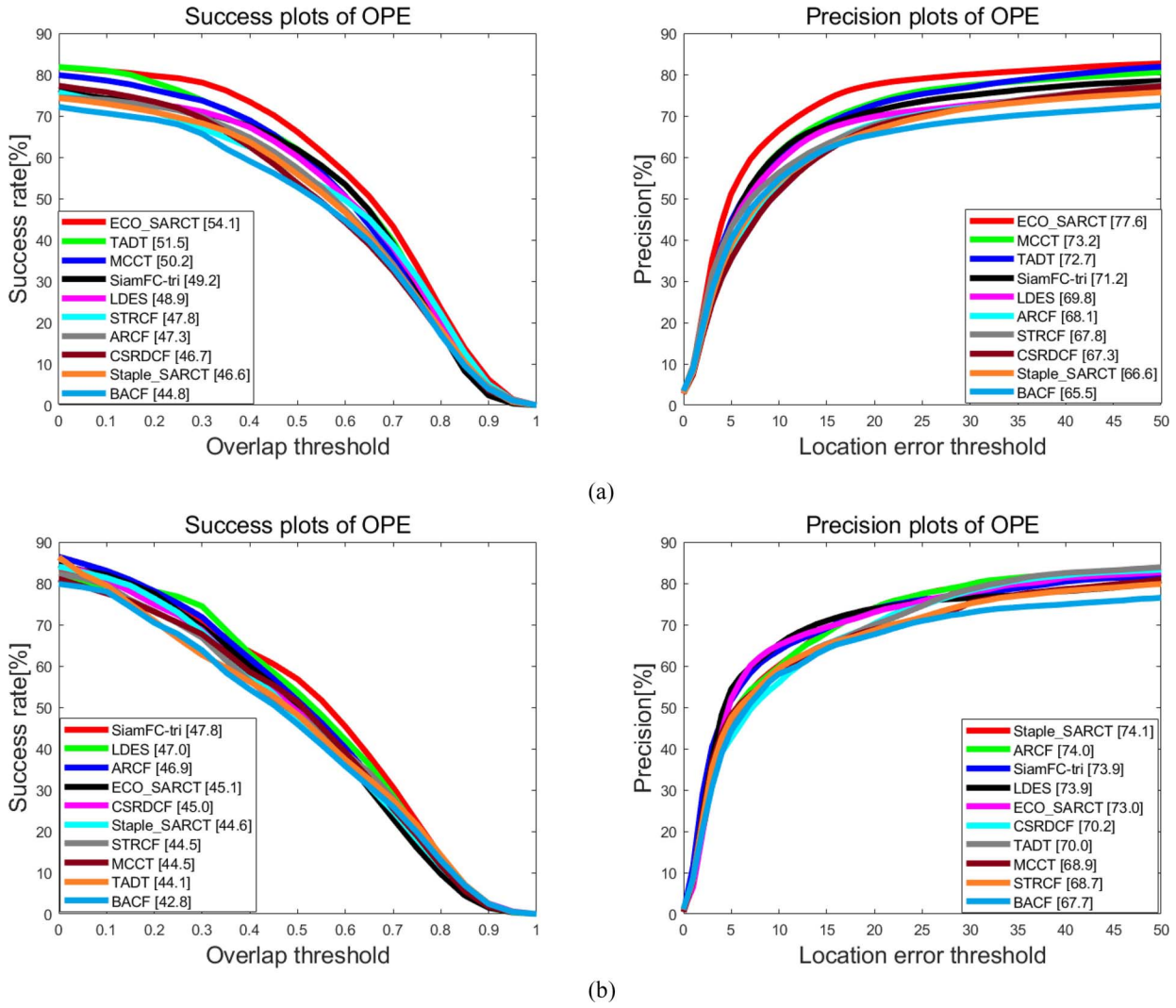


Fig. 6. Success and precision plots of two SARCT trackers and state-of-the-art methods on (a) UAV123 and (b) UAVDT datasets.

MB, low resolution, and out of view, expects the proposed ECO_SARCT. It is also observed that our ECO_SARCT has the ability to track the targets on sequences with full or partial occlusions (S0103) consistently. Specifically, ECO_SARCT manages to redetect the target while the other compared methods failed, including the aberrance repressed aerial tracker ARCF. Taking the advantage of semantic information, the proposed approach can locate the target even it has suffered from serious occlusion caused by the trees, and therefore outperform other state-of-the-art trackers. Although Staple_SARCT is also implemented under the proposed framework, it is not able to keep tracking the target throughout the sequence S0103, which proves that the selection of baseline affects the performance of the proposed framework to some extent. Implementing our framework on a more robust baseline may result in higher tracking accuracy.

E. Evaluation on Response Map Confidence

To a large extent, the peak fraction and the volatility level of the response map reflect the confidence level of the tracking outputs. If the tracking output perfectly matches to the real target location and scale, the desired response map should only own an obvious peak and should be slippery in all other regions. The more obvious the correlation peaks are, the better the location precision is. Otherwise, the response map might fluctuate violently [44]. To evaluate the performance from this aspect, a novel criterion called average peak to correlation energy (APCE) has been proposed in LMCF [44], defined by

$$\mathrm{APCE} = \frac{|F_{\max} - F_{\min}|^2}{\mathrm{mean}\left(\sum_{w,h}\left(F_{w,h} - F_{\min}\right)^2\right)}. \tag{12}$$

Following this work, the effect of the proposed SARCT framework is investigated regarding the APCE difference between the tracking performance of ECO_SARCT and that of ECO. From Fig. 7(a), it can be observed that although ECO and ECO_SARCT both successfully track the target during this period, the proposed approach provides a more accurate bounding box. Corresponding to this, the APCE of ECO_SARCT tracker on the sequence BMX4, shown in Fig. 8(a), has been significantly improved in comparison to the original ECO. This indicates that the response map resulting from the ECO_SARCT is of higher confidence.

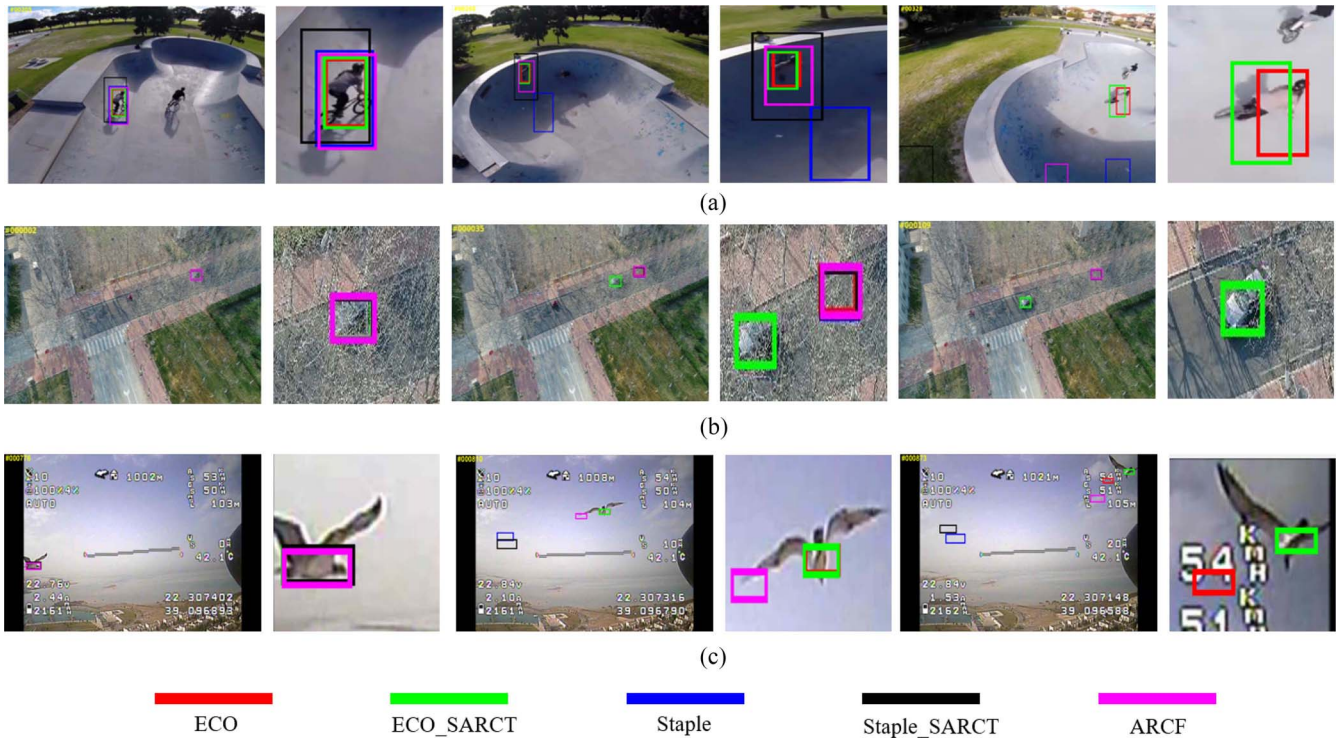
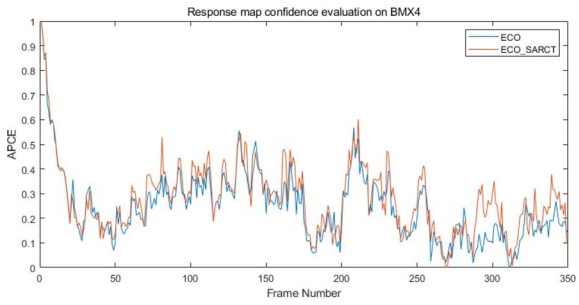
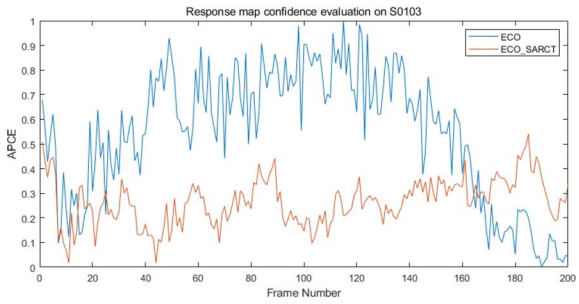


Fig. 7. Qualitative demonstration of the proposed ECO_SARCT, Staple_SARCT, their baselines, and the representative aerial tracker ARCF on sequences BMX4, S0103, and bird1 taken from DTB, UAVDT, and UAV123. (a) BMX4. (b) S0103. (c) bird1.



(a)



(b)

Fig. 8. Response map confidence evaluation of ECO_SARCT and ECO, compared on sequences (a) BMX4 and (b) S0103 taken from DTB and UAVDT datasets.

As shown in Figs. 7(b) and 8(b), when the target exhibits distinct appearance changes due to causes, such as sudden SV and partial or full occlusion, the proposed approach retains robust tracking while ECO loses target in the very beginning. As such, the APCE of ECO may be higher than ECO_SARCT in the first 160 frames, but it quickly drops down to a quite low value.

V. Conclusion

In this article, an SARCT framework has been proposed for UAV object tracking. SARCT is able to obtain accurate object locations with the assistance of an object detection procedure. By further introducing a semantic segmentation module to learn class-agnostic semantic information, SARCT adaptively reduces the background interference and alleviates the model drift problem suffered by the DCF baselines. Systematic evaluations have been carried out on three popular tracking benchmarks captured by UAVs. The results have demonstrated that SARCT achieves a significant improvement over various basic correlation filters while exhibiting at least equal performance to the state-of-the-art approaches, in terms of precision and success rate. Moreover, its speed can satisfy the real-time requirement for practical applications.

In the proposed framework, precise pixel-level semantic information is extracted from an image block under the consideration of computational burden. If the semantic information of an entire frame can be mined, abundant knowledge of the scenario categories and characteristics in the aerial scene of interest may be acquired. In so doing, instead of just computing the positions and tracks of a target, the target movement in aerial scenes may be better understood under a multitask framework consisting of semantic-aware object tracking and scene interpretation. Furthermore, the proposed SARCT may achieve performance enhancement by introducing more

advanced detection and segmentation techniques into its basic framework. The experimental verification of this conjecture remains active research.

REFERENCES

[1] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 375–391.

[2] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time UAV tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 2891–2900.

[3] X. Chen and Q. Meng, "Robust vehicle tracking and detection from UAVs," in *Proc. IEEE Int. Conf. Soft Comput. Pattern Recognit. (SoCPaR)*, Porto, Portugal, 2015, pp. 241–246.

[4] Z. Chen, Z. Hong, and D. Tao. (2015). *An Experimental Survey on Correlation Filter-Based Tracking*. [Online]. Available: https://arxiv.org/abs/1509.05520v1

[5] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 4282–4291.

[6] M. Danelljan, G. Bhat, F. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 4660–4669.

[7] Y. Wang, W. Shi, and S. Wu, "Robust UAV-based tracking using hybrid classifiers," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Venice, Italy, 2017, pp. 2129–2137.

[8] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2019.

[9] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 445–461.

[10] N. Wang, W. Zhou, and H. Li, "Reliable re-detection for long-term tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 730–743, Mar. 2019.

[11] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 21–37.

[12] M. Vondrak, L. Sigal, and O. C. Jenkins, "Dynamical simulation priors for human motion tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 52–65, Jan. 2013.

[13] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013.

[14] B. Wang, X. Yuan, X. Gao, X. Li, and D. Tao, "A hybrid level set with semantic shape constraint for object segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1558–1569, May 2019.

[15] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 2544–2550.

[16] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Apr. 2014.

[17] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, U.K., 2014, pp. 1–11.

[18] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1401–1409.

[19] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 4310–4318.

[20] C. Ma, J. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 3074–3082.

[21] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Santiago, Chile, 2015, pp. 58–66.

[22] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6638–6646.

[23] S. Li and D. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 4140–4146.

[24] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[25] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, 2014, pp. 188–203.

[26] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-term correlation tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 5388–5396.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3431–3440.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI.*, Munich, Germany, 2015, pp. 234–241.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[30] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1925–1934.

[31] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 801–818.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[35] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[36] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[37] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1369–1378.

[38] Y. Li, J. Zhu, S. C. Hoi, W. Song, Z. Wang, and H. Liu, "Robust estimation of similarity transformation for visual object tracking," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 8666–8673.

[39] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4844–4853.

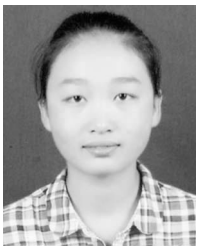
[40] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4904–4913.

[41] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 472–488.

[42] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6309–6318.

[43] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 1135–1143.

[44] M. Wang, Y. Liu, and Z. Huang, "Large margin object tracking with circulant feature maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4021–4029.

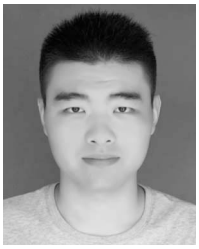


Xizhe Xue received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2018, where she is currently pursuing the Ph.D. degree with the School of Computer Science.

Her research interests include visual tracking, aerial image processing, and deep learning techniques.

Ying Li received the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2002.

Since 2003, she has been with the School of Computer Science, Northwestern Polytechnical University, Xi'an, where she is currently a Full Professor. Her current research interests include computational intelligence, image processing, and pattern recognition. She has published extensively in the above areas.



Xiaoyue Yin was born in 1997. He received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2019, where he is currently pursuing the master's degree in computer science.

His current research interests include computer vision and machine learning.

Changjing Shang received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1995.

She is a University Research Fellow with the Department of Computer Science, Aberystwyth University, Aberystwyth, U.K. She has published extensively, and supervised more than 15 Ph.D.s/Post-Docs in the areas of pattern recognition, data mining, space robotics, and image modeling and analysis.

Taoxin Peng received the Ph.D. degree in computer science from the University of Greenwich, London, U.K., in 2000.

He is a Lecturer with the School of Computing, Edinburgh Napier University, Edinburgh, U.K. His current research interests include data quality, data cleaning, data mining and data analytics, and model-based reasoning. His research findings have been published in both peer-reviewed international conferences and journals.

Qiang Shen received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1990, and the D.Sc. degree in computational intelligence from Aberystwyth University, Aberystwyth, U.K., in 2013.

He holds the established chair of computer science and is the Pro Vice-Chancellor for the Faculty of Business and Physical Sciences, Aberystwyth University. He has authored two research monographs and 400+ peer-reviewed papers.

Dr. Shen was a recipient of an Outstanding Transactions Paper Award from the IEEE.